

# Large-scale deployment of statistical machine translation Example Microsoft

Chris.Wendt@microsoft.com

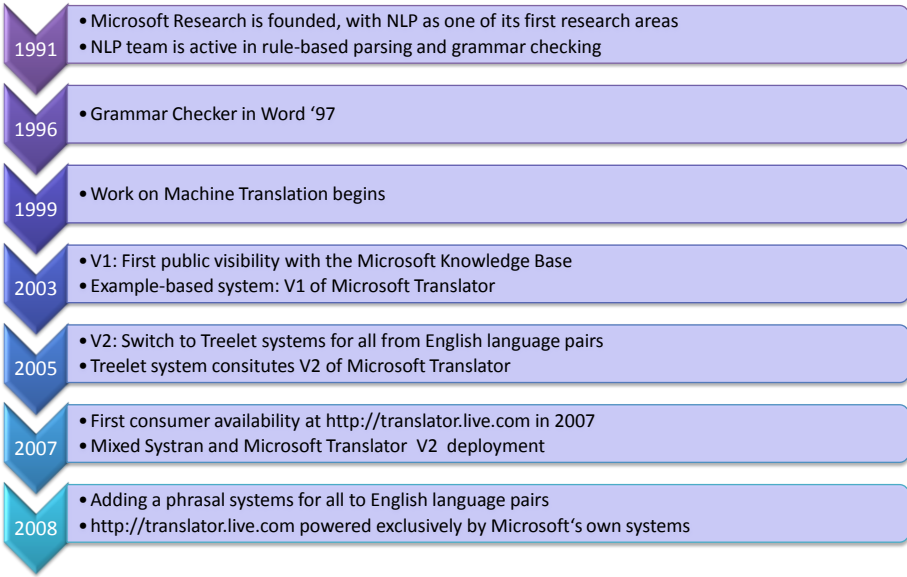
Microsoft Research – Machine Translation



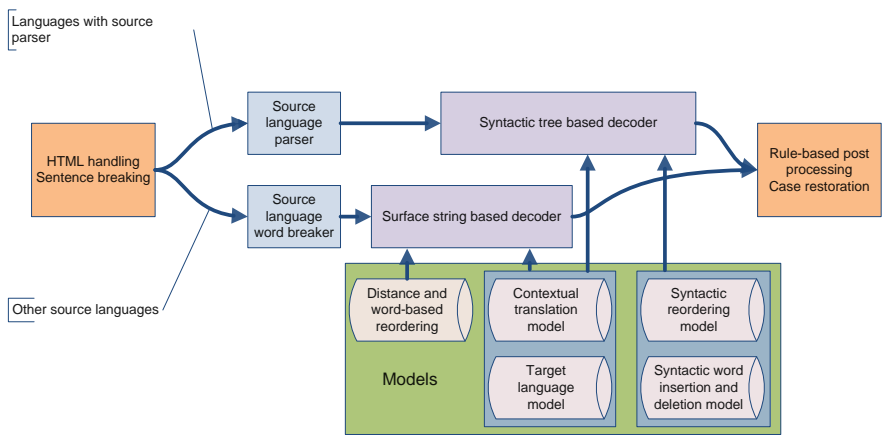
## Agenda

- Microsoft MT engine basics
- Architecture and design for scale
- Translator in Practice
- Microsoft internal use: Human Translation and Raw Publishing

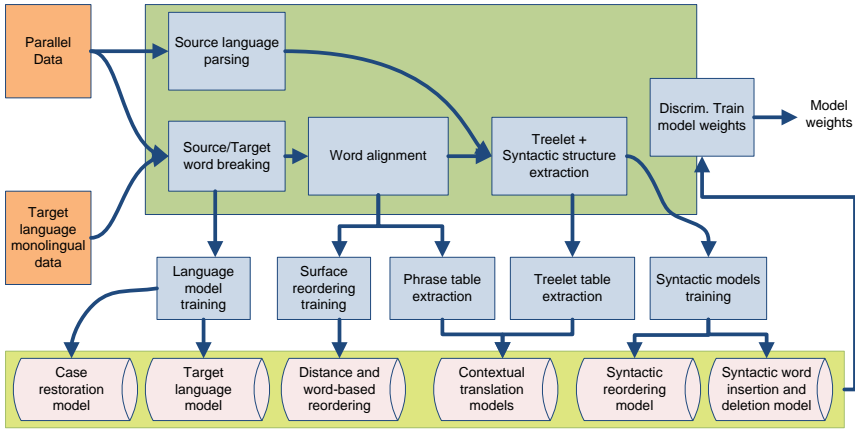
# Time Line



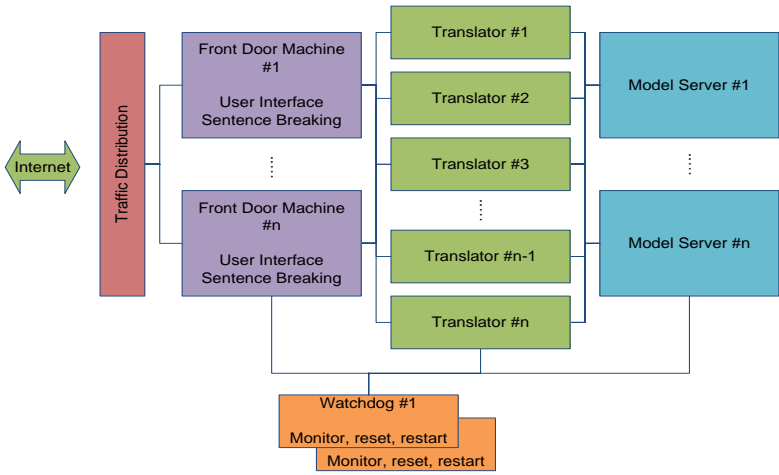
# Microsoft's Statistical MT Engine



# Training Architecture



# Runtime Architecture



## Front Door

- Microsoft Internet Information Server
- Landing Page
  - HTTP interface for Bilingual Viewer
    - Fetches web page, sentence & html breaking, creates marked up version
    - Sends page to client, asynchronously fills translation requests
- Distributor
  - SOAP API
  - Distributes sentences to multiple leaves
  - In memory cache of sentence translations

## Automatic evaluation: BLEU

- A fully automated MT evaluation metric
  - Modified N-gram precision, comparing a test sentence to reference sentences
- Automatic and cheap: runs daily and for every check-in
- Standard in the MT community
  - Immediate, simple to administer
  - Shown to correlate with human judgments
- Warning: Does not compare between engines or between languages.

# Human evaluations

- 3 to 5 independent human evaluators are asked to rank translation quality for 250 sentences on a scale of 1 to 4
  - Comparing to human translated sentence
  - No source language knowledge required

**4 = Ideal**  
Grammatically correct, all information included

**3 = Acceptable**  
Not perfect, but definitely comprehensible, *and* with accurate transfer of all important information

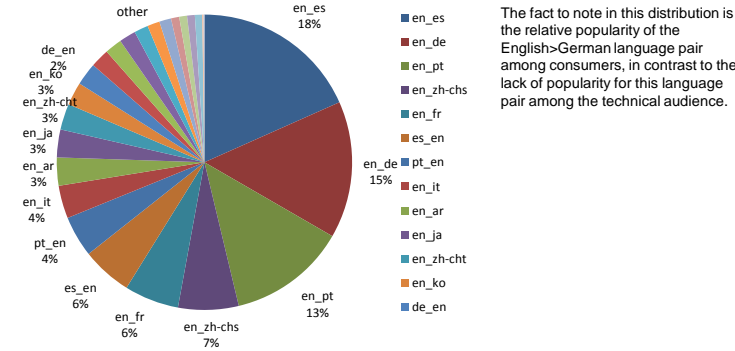
**2 = Possibly Acceptable**  
May be interpretable given context/time, some information transferred accurately

**1 = Unacceptable**  
Absolutely not comprehensible and/or little or not information transferred accurately

- Each sentence is evaluated by all raters, and scores are averaged
- Relative evaluations
  - Track progress against ourselves and a competitor

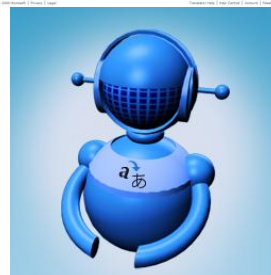
9

## Language pairs on translator.live.com



## Products

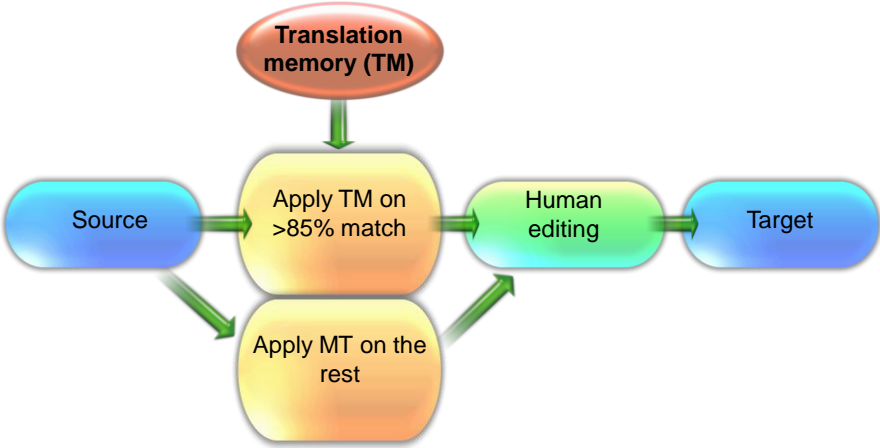
- Bilingual Viewer
  - Used by Live Search results page
- Translator landing page
- Toolbar Translator Button
- Translator Add-in for 3<sup>rd</sup> party pages
- Internet Explorer 8 accelerator
- Community built Firefox version
- Translator Bot (mtbot@hotmail.com)
- Office Research Pane
- SOAP API for product team use
- Microsoft Localization
  - CSS KB, MSDN Technet, Products



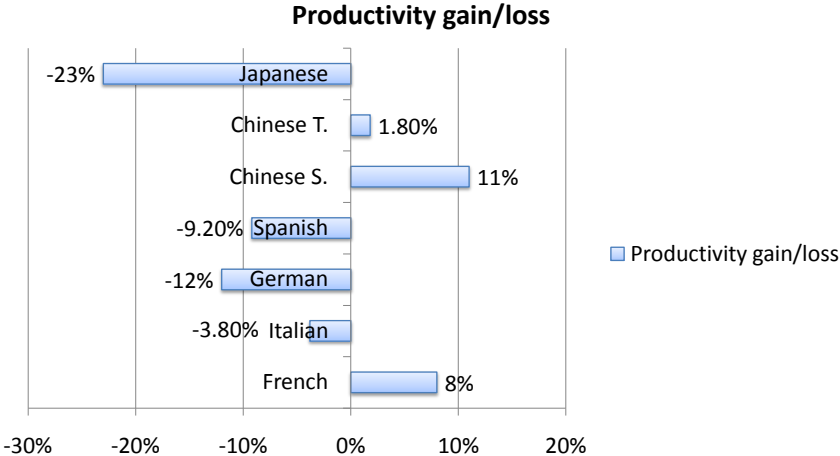
## Two ways to apply MT in a product

<ul style="list-style-type: none"> <li>• Post-Editing           <ul style="list-style-type: none"> <li>– Increase human translator's productivity</li> <li>– In practice: 0% to 25% productivity increase               <ul style="list-style-type: none"> <li>• Varies by content, style and language</li> </ul> </li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• Raw publishing           <ul style="list-style-type: none"> <li>– Publish the output of the MT system directly to end user</li> <li>– Best with bilingual UI</li> <li>– Good results with IT Pro and Developer audience</li> </ul> </li> </ul>
<p>→ Increasing the extent of localization</p>	

# MT with post-editing

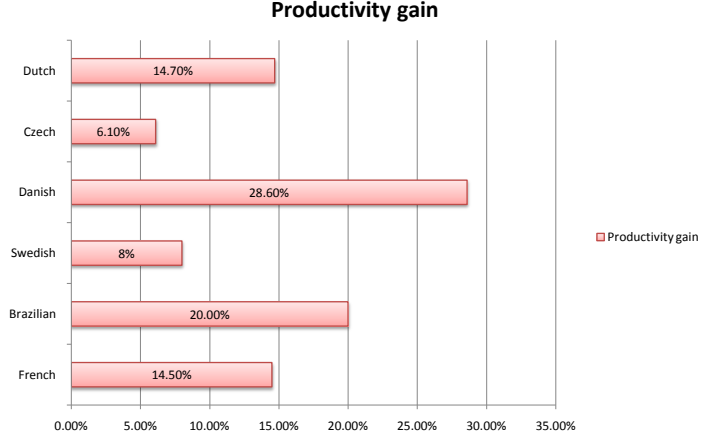


## Product 1: Post-editing Results Without specific post-editor training



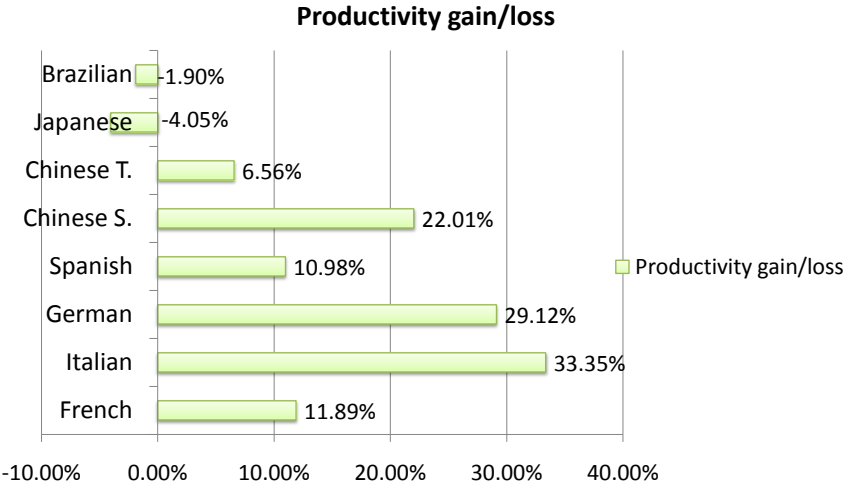
# Product 2: Post-Editing Results

## A couple of weeks later: with training



# Product 3: Post-editing Results

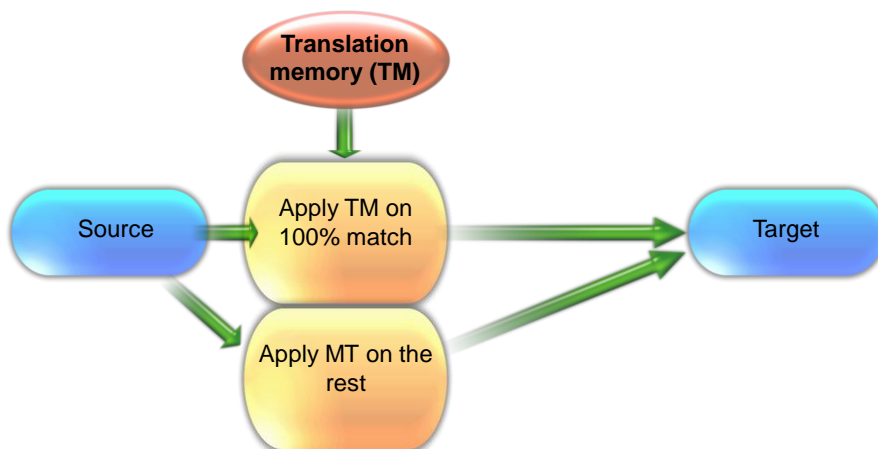
## With training



## Post-Editing: Lessons Learned

- Training of the translator is required
  - Understand the peculiarities of the engine used
  - Always read the source sentence first
  - Understand when to discard the MT
    - “Two seconds is Too much”
- Acknowledge different suitability for different style and terminology
- Customize terminology per individual project – use of dictionary
- Productivity gains of 5% to 25% are achievable, but investment is required

## Raw MT Publishing



España Enlaces Inicio Otros países

**Microsoft**

Ayuda y Soporte Técnico

Página de inicio Seleccione un producto Buscar Grupos de noticias Contáctenos Mapa del sitio

## Búsqueda de Ayuda y Soporte Técnico

Usted buscó [Obtener Ayuda](#) | [Búsqueda Avanzada](#)

Buscar en:
 

- Knowledge Base en Español
- Knowledge Base en inglés
- Todo Microsoft.com
- Internet

Producto: Todos los Productos

Buscar por: novell netware

Herramientas
 

- [Imprimir la página](#)
- [Enviar esta página](#)
- [Microsoft en el mundo](#)
- [Mis favoritos de soporte](#)
- [Iniciar sesión](#)

Esta buscando: [novell netware](#)

Resultados de la búsqueda

Mostrar resultados para:
 

- [Systems Management Server](#)
- [Windows 95](#)
- [Windows 95](#)
- [Windows 2000 Professional Edition](#)
- [Windows 2000](#)
- [Systems Management Server 2.0](#)
- [Word 2000](#)

Resultados 1-20 de 200+ [Siguiente >](#) [Mostrar todos](#)

- \* [Lista de los privilegios que se requieren en una red Novell NetWare para ejecutar Word](#) (244710) - Describe los privilegios que se necesitan para realizar operaciones específicas Word en una red Novell NetWare. También describe los síntomas de privilegios faltados. <http://support.microsoft.com/kb/244710/es>
- \* [Cómo migrar o implementar un entorno Novell NetWare en Windows](#) (274279) - Si está considerando migrar a Windows de Novell NetWare o si necesita información acerca de una implementación de entorno mixto, este artículo proporciona estrategias y recomendaciones de obtener de iniciar.

Internet | Protected Mode: On 100%

España Enlaces Inicio Otros países

Microsoft

Ayuda y Soporte Técnico

Página de inicio Seleccione un producto Buscar Grupos de noticias Contáctenos Mapa del sitio

## Lista de los privilegios que se requieren en una red Novell NetWare para ejecutar Word

[Ver los productos a los que se aplica este artículo](#)

**AVISO:** Gracias por utilizar el servicio de Traducción Automática. Este artículo ha sido traducido por un sistema informático sin ayuda humana (Machine Translation). Microsoft ofrece estos artículos a los usuarios que no comprendan el inglés, exclusivamente, con el fin de que puedan entenderlos más fácilmente. Microsoft no se hace responsable de la calidad lingüística de las traducciones ni de la calidad técnica de los contenidos de los artículos así como tampoco de cualesquiera problemas, directos o indirectos, que pudieran surgir como consecuencia de su utilización por los lectores.

En esta página
 

- [Resumen](#)
- [Más información](#)
- [Descripción de directorio de sistema de archivos y privilegios de archivo](#)
- [Privilegios de Novell que requieren sólo leer documentos](#)
- [Privilegios de Novell que requieren crear o modificar documentos](#)
- [Síntomas de perder privilegios de Novell](#)
- [Por qué Novell incluye Digitalizar de Archivo así como crea privilegios](#)

Id. de artículo : 244710
Última revisión : jueves, 25 de enero de 2007
Versión : 4.0

**Resumen**

Los privilegios de red que tienen que tener que realizar operaciones específicas de Microsoft Word por encima de unas redes Novell en este artículo se enumeran usted. El Este artículo incluye las secciones siguientes:

- Descripción de directorio de sistema de archivos y privilegios de archivo
- Privilegios de Novell que requieren sólo leer documentos
- Privilegios de Novell que son necesarios para crear o modificar documentos
- Síntomas de perder privilegios de Novell
- Por qué Novell incluye Digitalizar de Archivo así como crea privilegios

[Volver al principio](#)

**Más información**

Buscar en: novell netware

Knowledge Base en Esp. [Búsqueda Avanzada](#)

Seleccione idioma: Alemán

**Centros de Soluciones relacionados**

- [Word 2007](#)
- [Word 2003](#)
- [Word 2002](#)
- [Word 2000](#)

**Más soporte técnico**

- [Contactar con Microsoft](#)  
Opciones de Soporte, Teléfonos de Soporte, Soporte Online
- [Presentaciones Técnicas](#)  
Presentaciones Técnicas de nuestros productos ofrecidas por expertos de Microsoft.
- [Grupos de noticias](#)  
Exponga su consulta, problemas y que otros usuarios le propongan soluciones.

Herramientas
 

- [Imprimir la página](#)
- [Enviar esta página](#)

Internet | Protected Mode: On 100%

## History of MT in Customer Support

- Since 2003 CSS has been actively using Machine Translation for Knowledge Base articles
  - Spanish was the first language deployed
  - Japanese went live one year later
- Current Languages
  - 10 Languages deployed: Spanish, German, French, Italian, Japanese, Portuguese, Brazilian Portuguese, Chinese Simplified, Chinese Traditional, Arabic
  - 3 Languages in Testing: Korean, Turkish and Russian

## Microsoft Knowledge Base

### MT & HT distribution across languages

Language	Human translated, or originally authored in language	%
English	235,425	<b>100%</b>
Japanese	70,684	<b>27%</b>
French	35,310	<b>14%</b>
German	30,459	<b>12%</b>
Spanish	16,980	<b>7%</b>
Italian	14,401	<b>6%</b>
Chinese (Simplified)	12,873	<b>5%</b>
Chinese (Traditional)	10,372	<b>4%</b>
Portuguese (Brazil)	10,205	<b>4%</b>
Portuguese (Iberian)	7,129	<b>3%</b>
Arabic	2,152	<b>1%</b>

Traffic to the knowledge base is fairly unevenly distributed. By targeting human translation to the high page view articles, 80% of the Japanese total page views are for human translated articles. Even in Arabic 54% of the page views end up on human quality articles.

# Customer Feedback: KB Inline Survey

**Provide feedback on this article**

Did this article help you solve your problem?

Yes  
 No  
 Partially  
 I do not know yet

Strongly Agree 9 8 7 6 5 4 3 2 1 Strongly Disagree

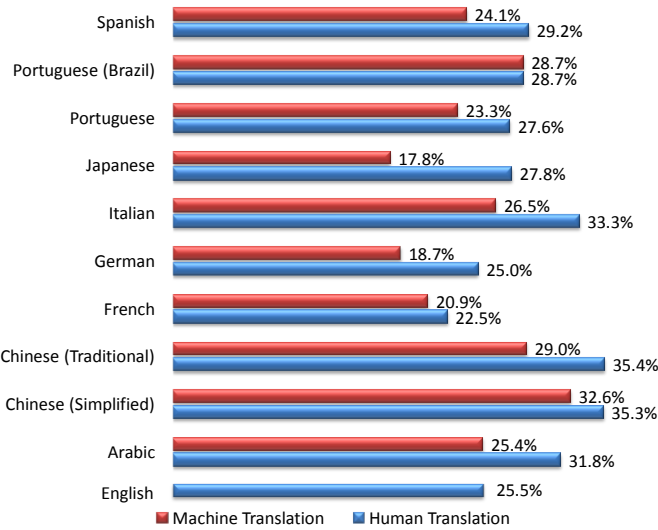
The article is easy to understand

The article is accurate

**Additional Comments:**

To protect your privacy, do not include contact information in your feedback.

## Knowledge Base – average resolve rate of human translated vs. machine translated articles

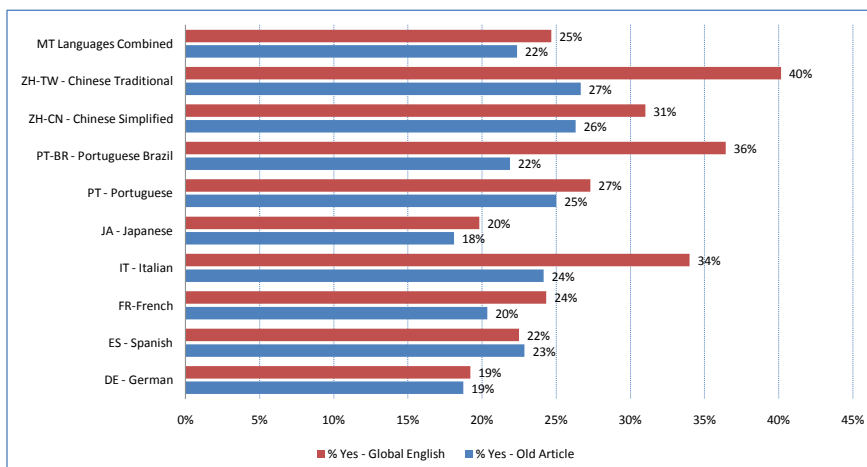


## Global English

- Support started rewriting source language to account for MT in 2003 (6 months after Spanish MT was deployed)
- Retrained the writers to write with global audience and MT in mind.
- Top five rules to make source language content suitable for MT:
  1. Use Standard English writing style
  2. Use correct punctuation – especially the following:
    - Missing punctuation causing incorrect sentence break
    - Hyphens
    - Commas
  3. Eliminate long sentences
  4. Use capitalization correctly
  5. Use correct spelling

## Impact of Global English

Resolve rate of articles authored to standard guidelines



# Translation Wiki

msdn  
Pesquisar MSDN com a Live Search

Visual Studio 2008  
Página inicial | Biblioteca | Aprenda | Downloads | Suporte | Comunidade | Informações sobre produtos

Versão Amigável para Impressora | Adicionar a Favoritos | Enviar | Adicionar conteúdo...

Ferramentas de desenvolvimento e linguagens

- Visual Studio 2008
  - Visual Studio
    - Instalação e configuração conceitos básicos
    - Introdução ao Visual Studio
    - Desenvolvimento de aplicativos em Visual Stu
    - Aplicativos com base no Windows, Component
    - Programação no Visual Studio .NET Framework
    - Visual Basic
    - Visual C#
    - Visual C++
    - Visual J#
    - JScript
    - Visual Web Developer
    - Documentação do Visual Studio Tools for Offic
    - Smart Device Development
    - Ferramentas e recursos
    - Visual Studio e .NET Framework Glossário

Visual Studio

Recoher Tudo | Exibir Conteúdo: Lado a Lado

Este conteúdo foi traduzido automaticamente e pode ser editado pelos membros da comunidade. Para melhorar a qualidade da tradução, clique no link Editar associado à frase que deseja modificar.

Visual Studio

**Visual Studio**

Visual Studio is a complete suite of tools for building both desktop and team-based enterprise Web applications. In addition to building high-performing desktop applications, you can use Visual Studio's powerful component-based development tools and other technologies to simplify team-based design, development, and deployment of enterprise solutions.

In This Section

Visual Studio

**Visual Studio**

Visual Studio é um conjunto completo de ferramentas para criar aplicativos para desktops e aplicativos Web empresariais baseados em equipes. Além de criar aplicativos individuais de alto desempenho, você pode usar as poderosas ferramentas de desenvolvimento baseadas em componentes do Visual Studio e outras tecnologias para simplificar o projeto, desenvolvimento e implantação de soluções empresariais baseadas em equipe.

## Benefits of using MT

Larger language set

- Localize into more languages without increasing budget

Localize More

- Increase the extent of localization without proportional budget increase

Higher Productivity

- 5% to 25% productivity increase
- >25% in software localization

Faster Availability

- Remove delay in translation
- Especially desired by technical audience

## Conclusions

- Automation enriches the customer's experience
- Long term investments in tools and processes are required, and patience in seeing results is needed
- Reaching customers through multiple forums and media is important
- Metrics are more useful than opinions
- Using customer feedback and community provides better solutions

Thank you

# References

- Menezes, Arul, Kristina Toutanova and Chris Quirk. Microsoft Research Treelet translation system: NAACL 2006 Europarl evaluation. Workshop on Machine Translation, NAACL 2006
- Chris Quirk, Arul Menezes. Dependency Treelet Translation: The convergence of statistical and example-based machine translation? March 2006 Machine Translation 43--65